ETH zürich

Longitudinal early warning systems

Thomas Gumbsch

May 2022

Patient monitoring

Patient monitors produce alarms if clinical variables reach predefined thresholds. These are

a) unspecific and
b) not individualized

causing *alarm fatigue*: Clinicians get desensitized towards alerts so much that those are turned off out of routine.

Potential for a machine learning solution.

ETH zürich

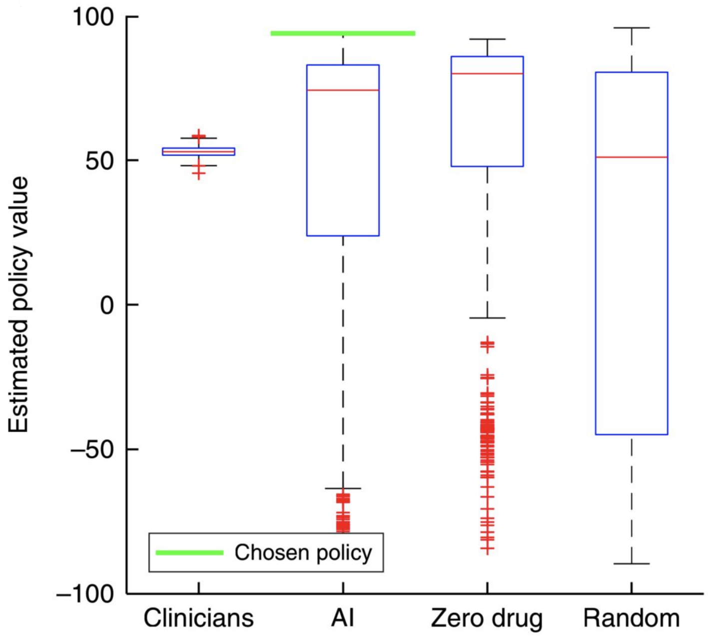
Trust issue in the healthcare domain

ARTICLES
https://doi.org/10.1038/s41591-018-0213-5
nature medicine

The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care

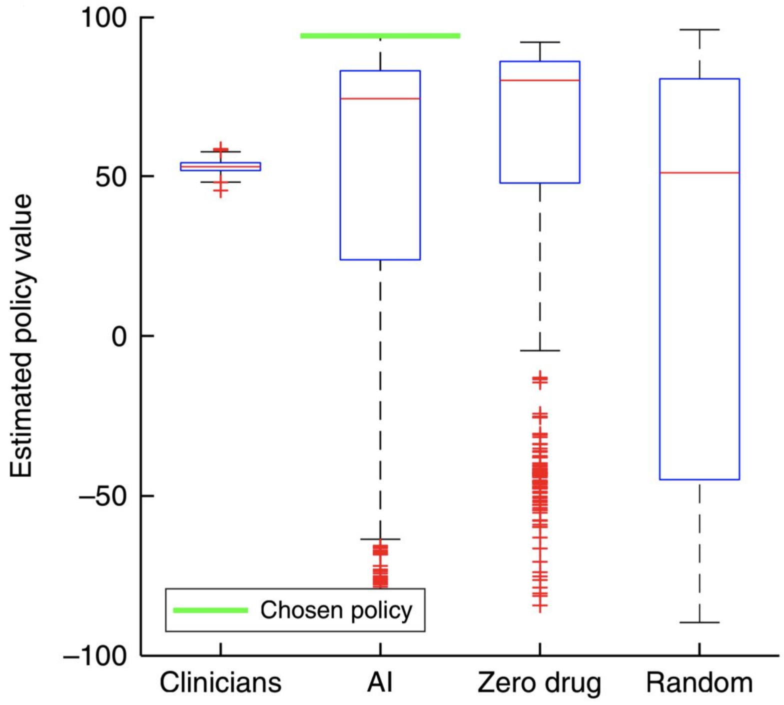
Matthieu Komorowski[1,2,3], Leo A. Celi[3,4], Omar Badawi[3,5,6], Anthony C. Gordon[1]* and A. Aldo Faisal[2,7,8,9]*

Sepsis is the third leading cause of death worldwide and the main cause of mortality in hospitals[1-3], but the best treatment strategy remains uncertain. In particular, evidence suggests that current practices in the administration of intravenous fluids and vasopressors are suboptimal and likely induce harm in a proportion of patients[1,4-6]. To tackle this sequential decision-making problem, we developed a reinforcement learning agent, the Artificial Intelligence (AI) Clinician, which extracted implicit knowledge from an amount of patient data that exceeds by many-fold the life-time experience of human clinicians and learned optimal treatment by analyzing a myriad of (mostly suboptimal) treatment decisions. We demonstrate that the value of the AI Clinician's selected treatment is on average reliably higher than human clinicians. In a large validation cohort independent of the training data, mortality was lowest in patients for whom clinicians' actual doses matched the AI decisions. Our model provides individualized and clinically interpretable treatment decisions for sepsis that could improve patient outcomes.



Why has the zero drug policy a higher expected value than the clinicians policy?

Trust issue in the healthcare domain



Does the "Artificial Intelligence Clinician" learn optimal treatment strategies for sepsis in intensive care?

Russell Jeter*[1], Christopher Josef*[2], Supreeth Shashikumar[3], and Shamim Nemati[1]
[1]Department of Biomedical Informatics, Emory University, Atlanta, USA.
[2]School of Medicine, Department of Surgery, Emory University, Atlanta, USA.
[3] School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, USA.

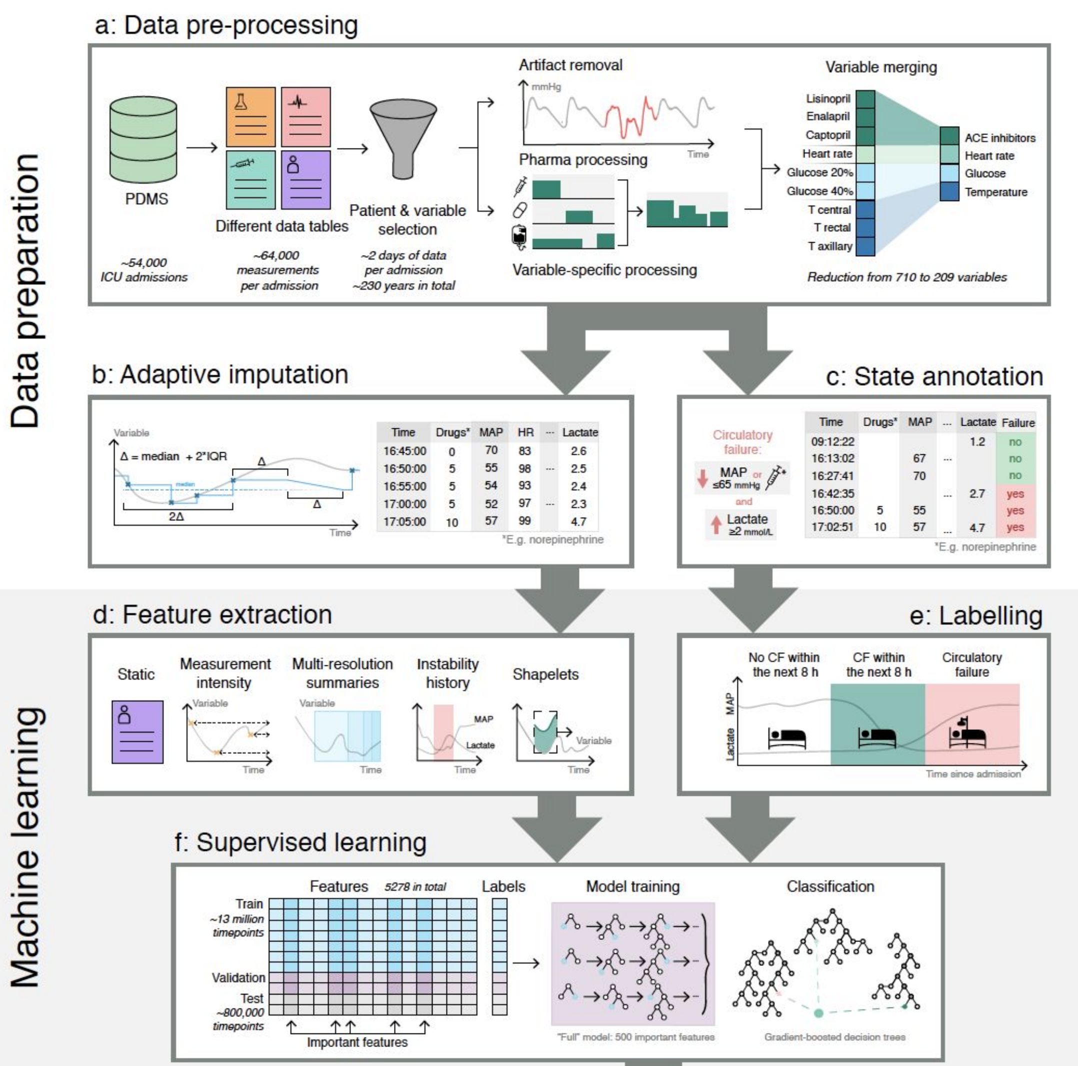
1) Importance sampling gives high weight to patients that are stable and not treated
2) AI learns to game the system by acting differently than clinician in tough cases

Primum non nocere (first do no harm)

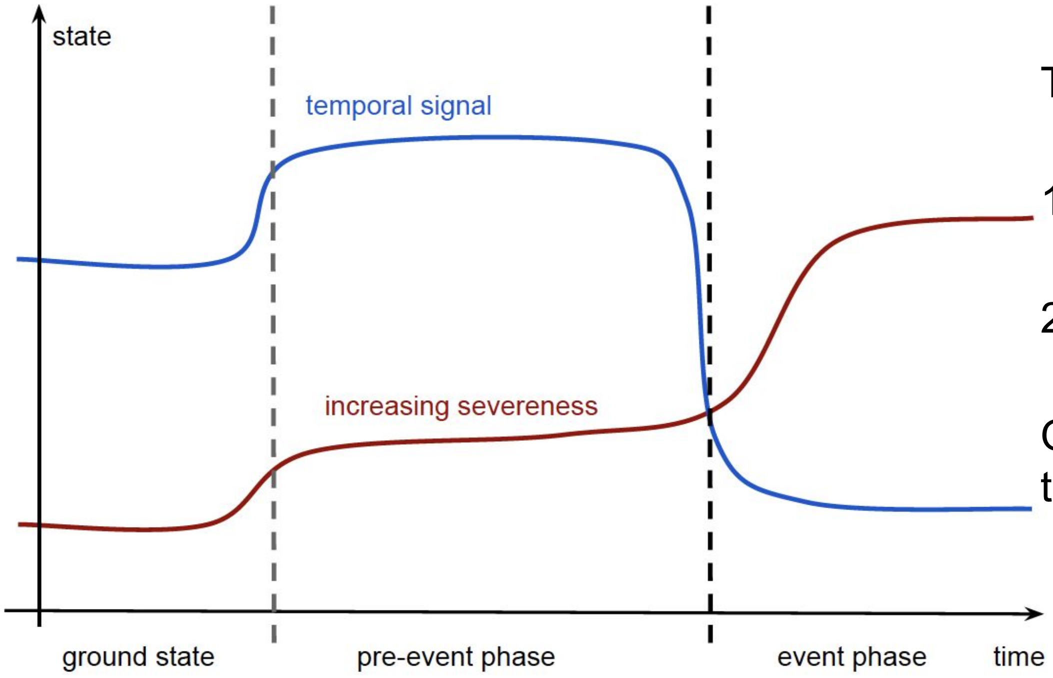
Part 1: The setup

Build a machine learning early warning system …

1. … for event prognosis
2. … of circulation failure
3. … in the intensive care unit in Bern, Switzerland.



Machine learning prognosis systems



Train a binary classifier on timepoints

1. detecting deviations towards an event

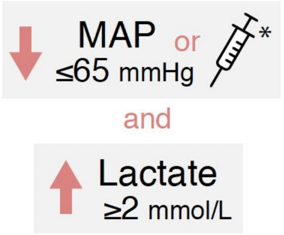
2. **detecting a pre-event phase.**

One sample is a feature vector per timepoint on a 5 min grid.

Circulation failure

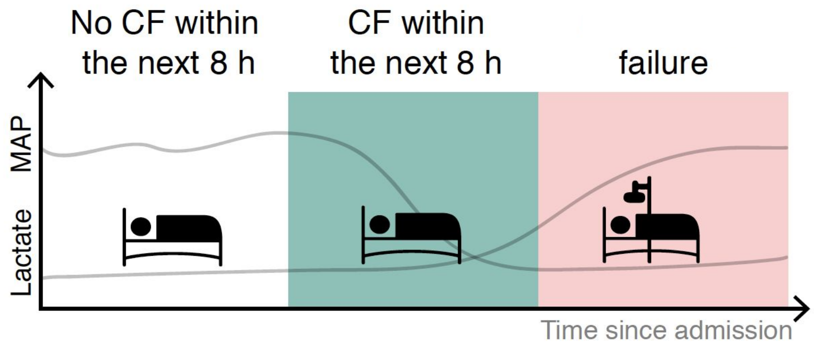
Bersten, A. D., & Handy, J. (2013). *Oh's Intensive Care Manual Chapter 92*. Elsevier Health Sciences.

Detecting a sustained period of low mean arterial blood pressure or receiving vasopressors with elevated serum lactate...



Time	Drugs*	MAP	...	Lactate	Failure
09:12:22				1.2	no
16:13:02		67	...		no
16:27:41		70			no
16:42:35			...	2.7	yes
16:50:00	5	55			yes
17:02:51	10	57	...	4.7	yes

*E.g. norepinephrine



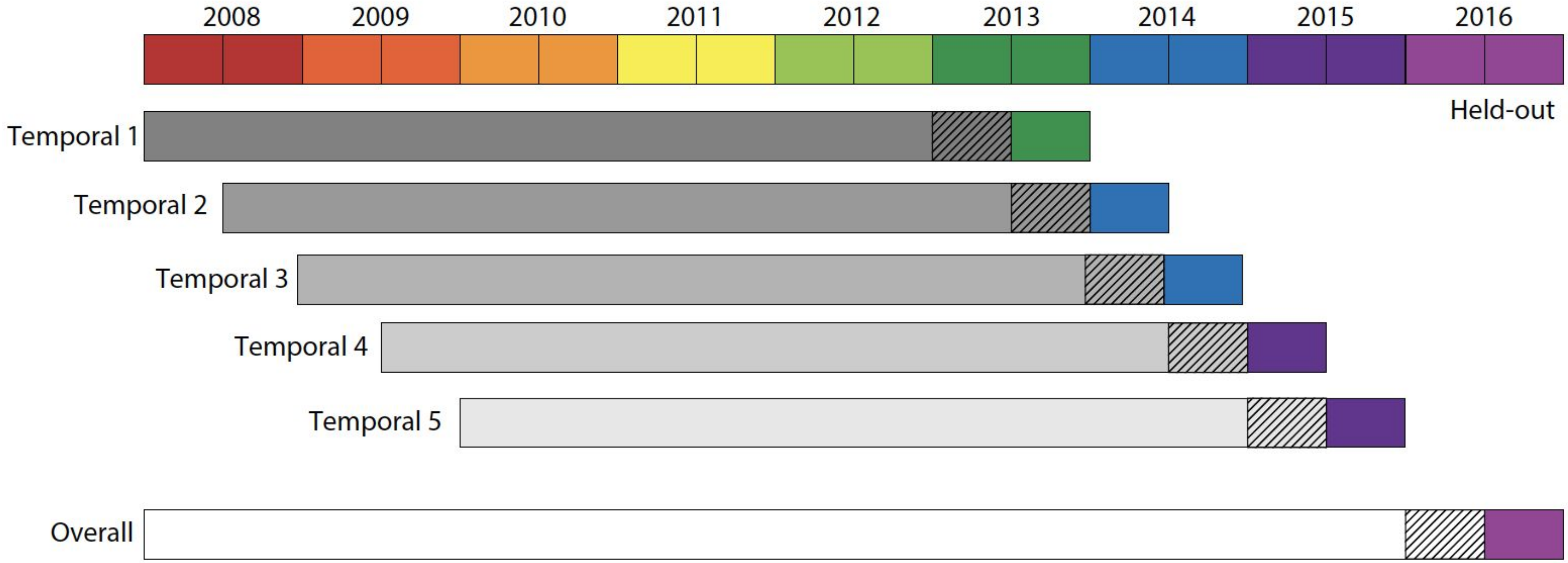
...until the next shift, i.e. 8 hours.

ICU databases

	MIMIC-III	eICU	HiRID	MIMIC-IV
Stays	60k	200k	55k	70k
Frequency	60m	5m	3m	60m
Variables	31′800	23′500	7′300	5′500
Observations	300m	2′800m	3′000m	350m
Timespan	2001 – 2012	2014 – 2015	2008 – 2015	2008 – 2019
Median LOS	2.10d	1.57d	0.93d	2.06d
Units	5	335	1	1
Benchmarking	(Harutyunyan et al., 2019)	no	(Yèche et al., 2021)	no
non-ICU	no	no	(yes)	yes

HiRID gives 7.5 observations per patient and per variable key, in comparison to 0.1/0.6/0.9 for MIMIC-III / eICU / MIMIC-IV legitimating the name of a high-resolution ICU database (HiRID).

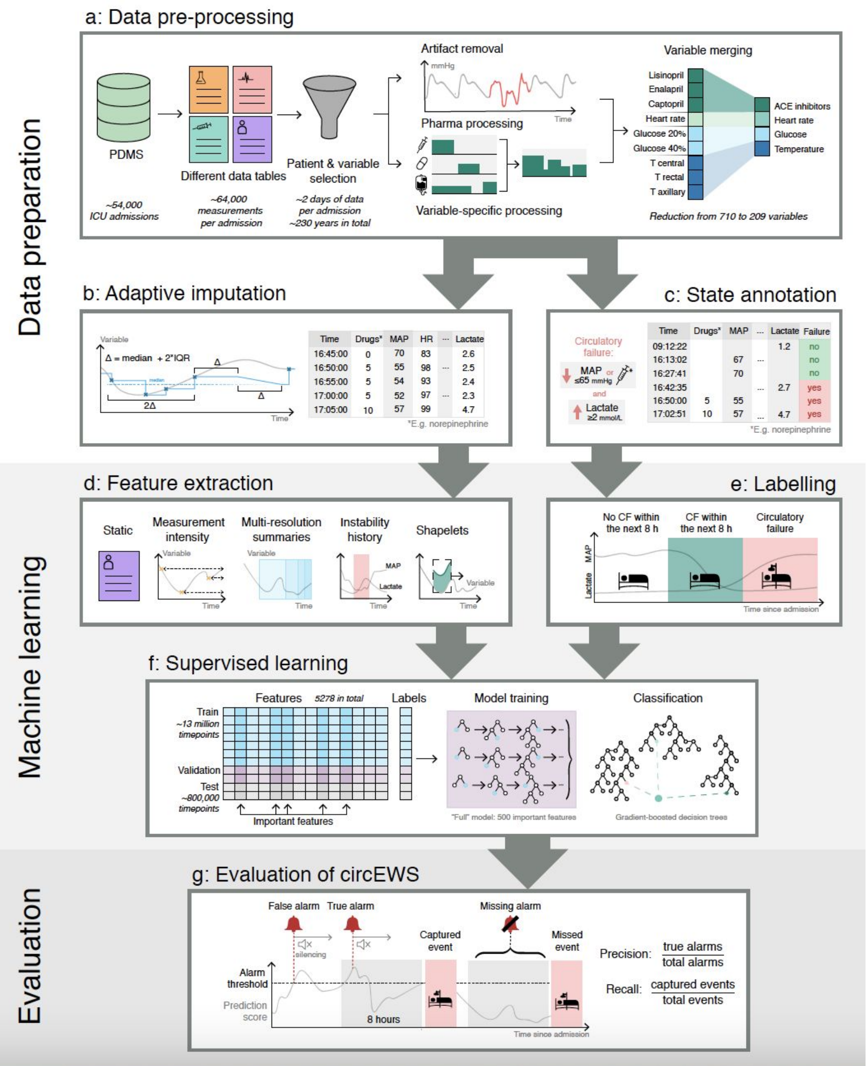
Estimating model variance



ETH zürich

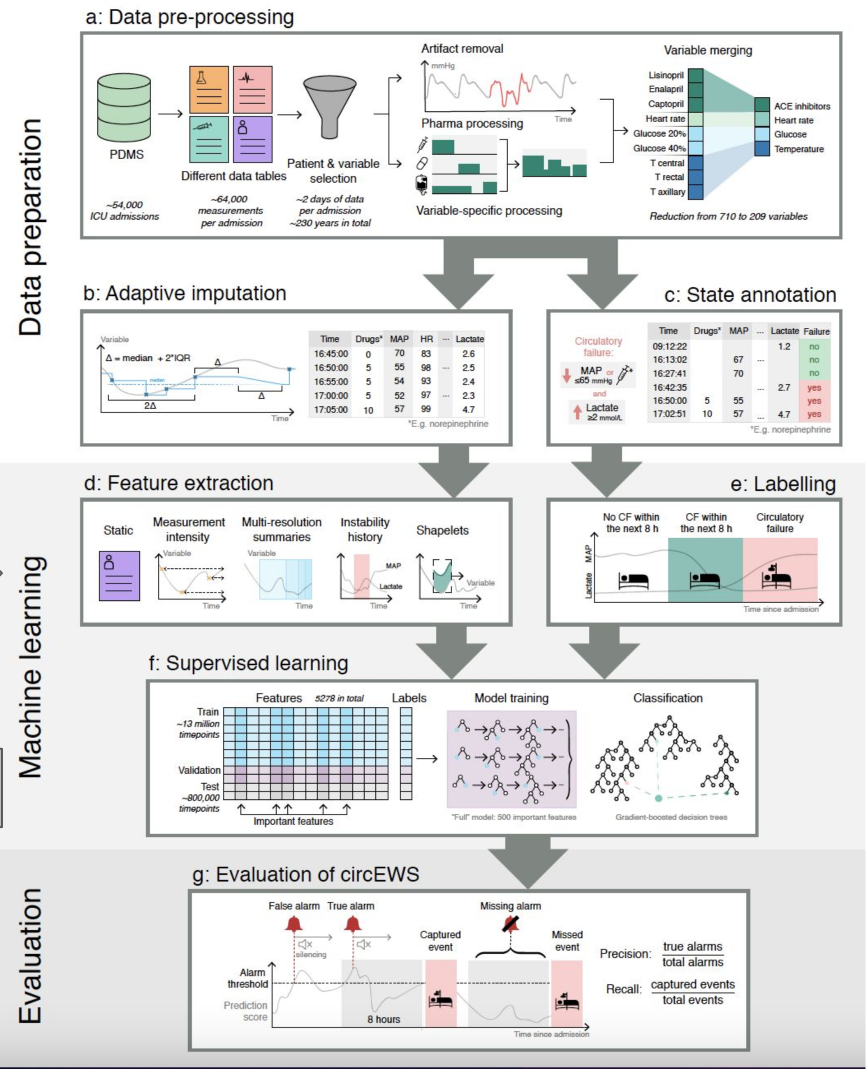
Methods overview

- **circEWS** (circulatory early warning system) derived from the predictions of a random forest on all features of HiRID
- circEWS-lite reduced model
- Baseline decision tree on the endpoint-defining variables
- MEWS, a severity score



Overview

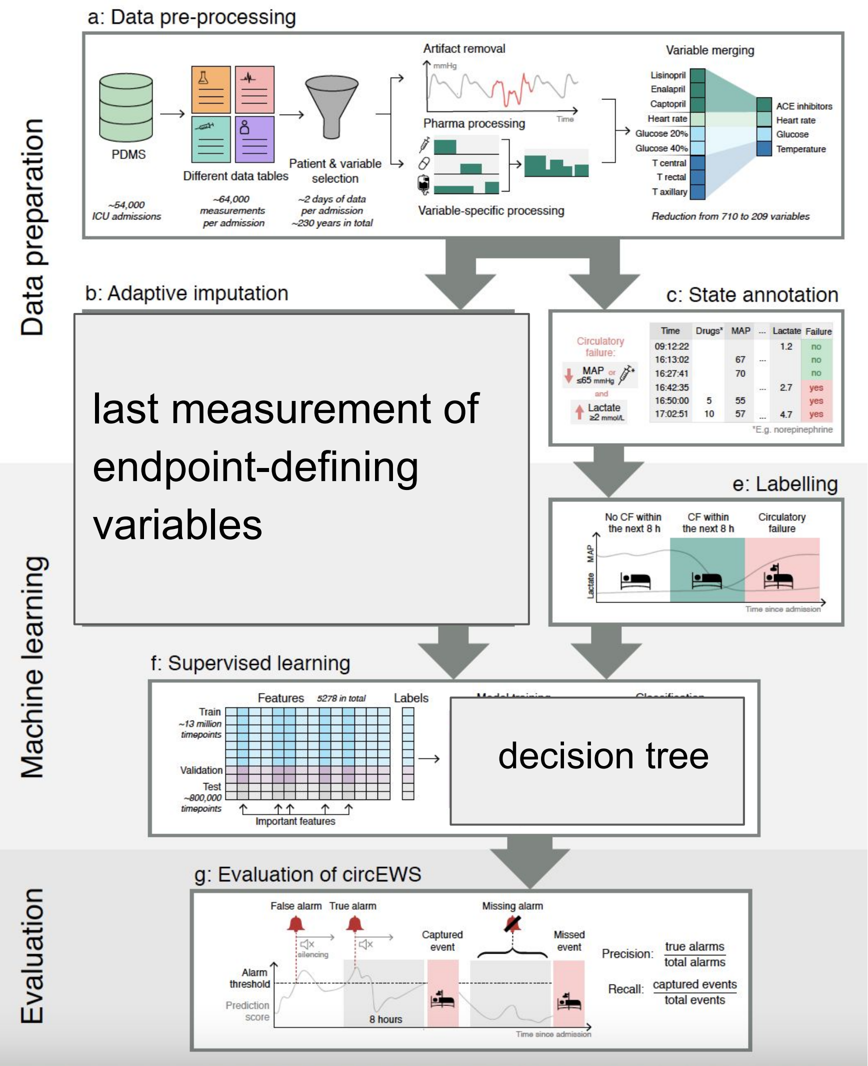
- circEWS full model
- **circEWS-lite** reduced model using only 20 most important variables of circEWS on HiRID
- Baseline decision tree on the endpoint-defining variables
- MEWS, a severity score



ETH zürich

Overview

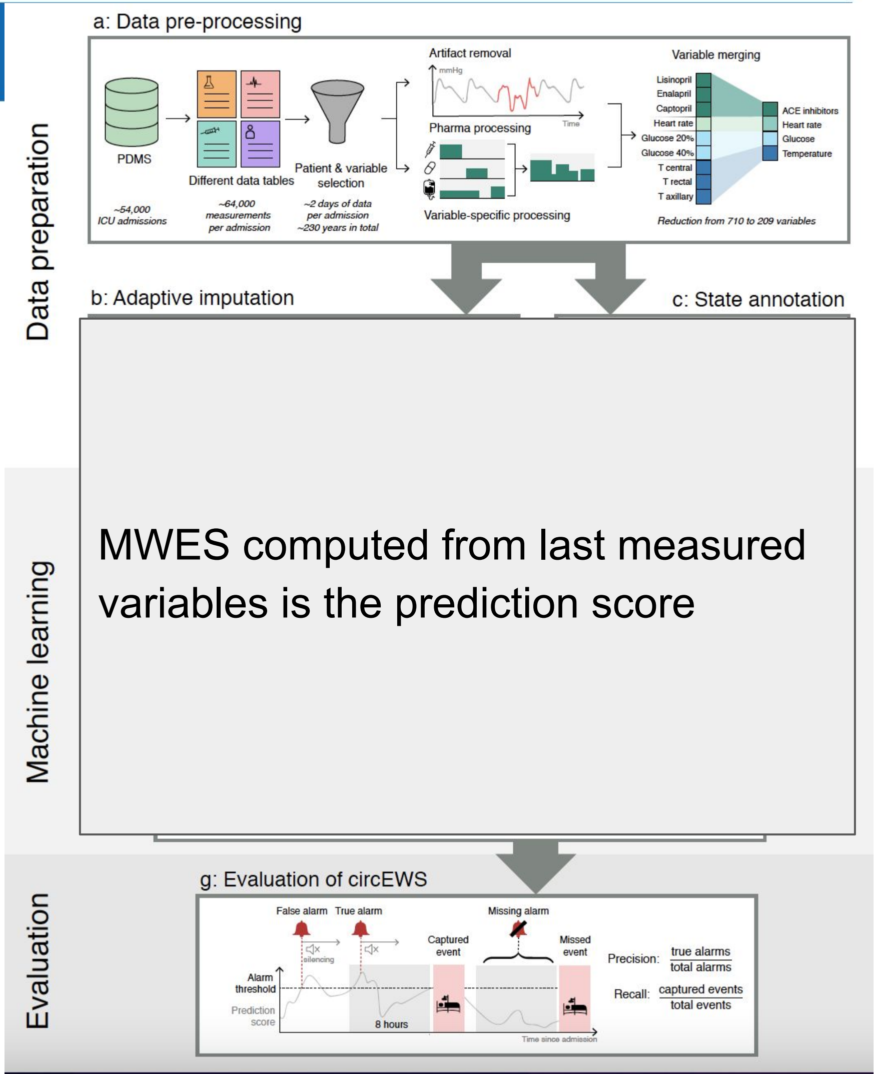
- circEWS full model
- circEWS-lite reduced model
- **Baseline** decision tree on the endpoint-defining variables
- MEWS, a severity score



ETH zürich

Overview

- circEWS full model
- circEWS-lite reduced model
- Baseline decision tree on the endpoint-defining variables
- **MEWS** is an early warning system by treating the severity score as a prediction.



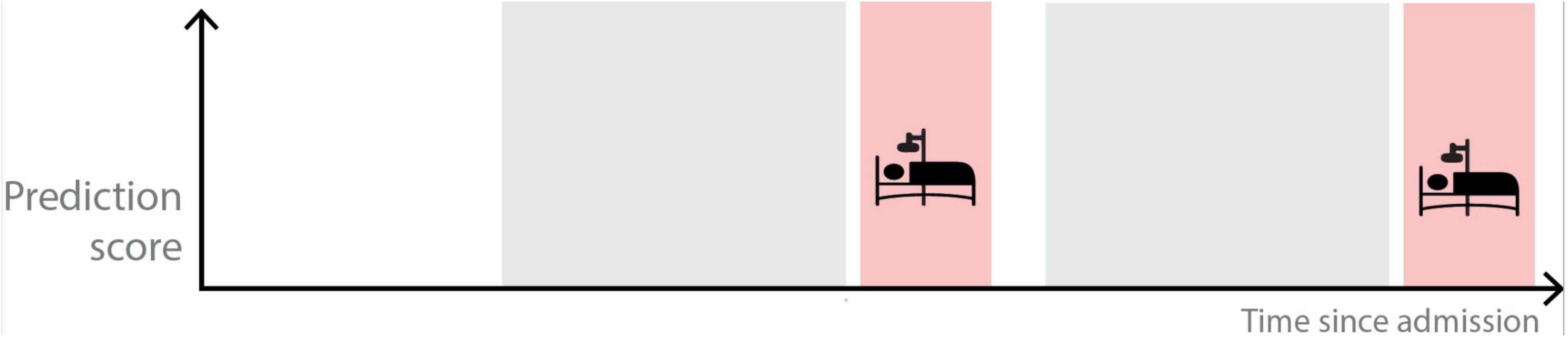
Part 2: The evaluation

Evaluate methods for task of circulation failure prediction with different measures

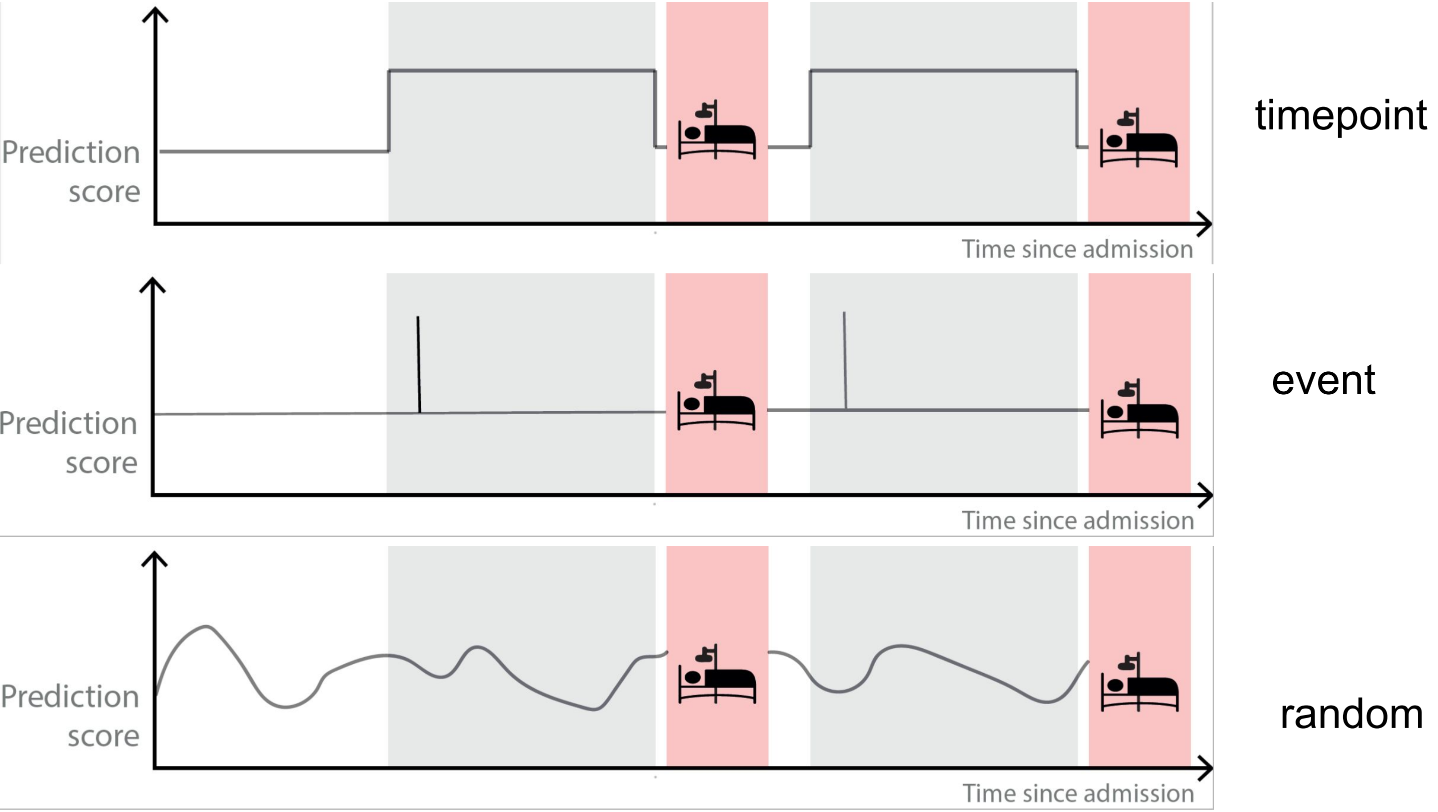
1. **Timepoint-based classification evaluation**
2. **Event-based binary classification evaluation**
3. **Maintenance policy evaluation**
4. **Control chart methods**

ETH zürich

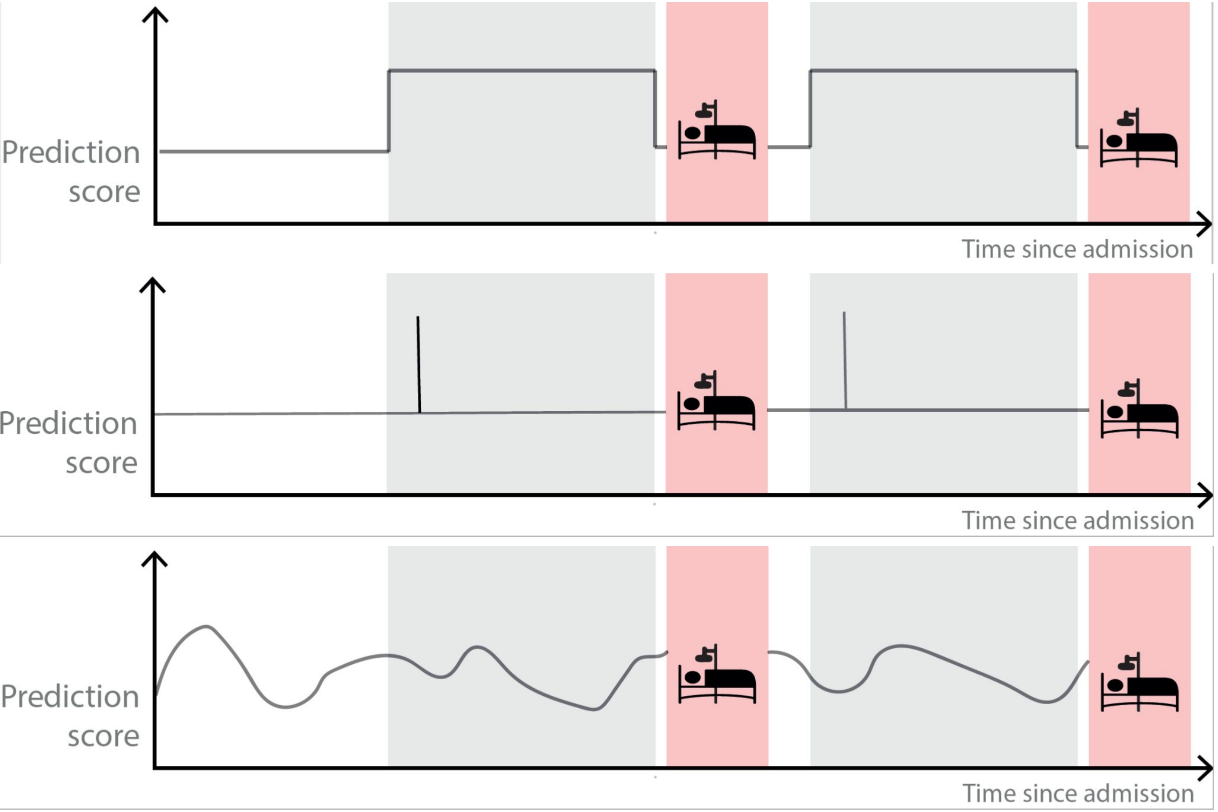
Binary classification evaluation



Synthetic comparison partners



Synthetic comparison partners



timepoint

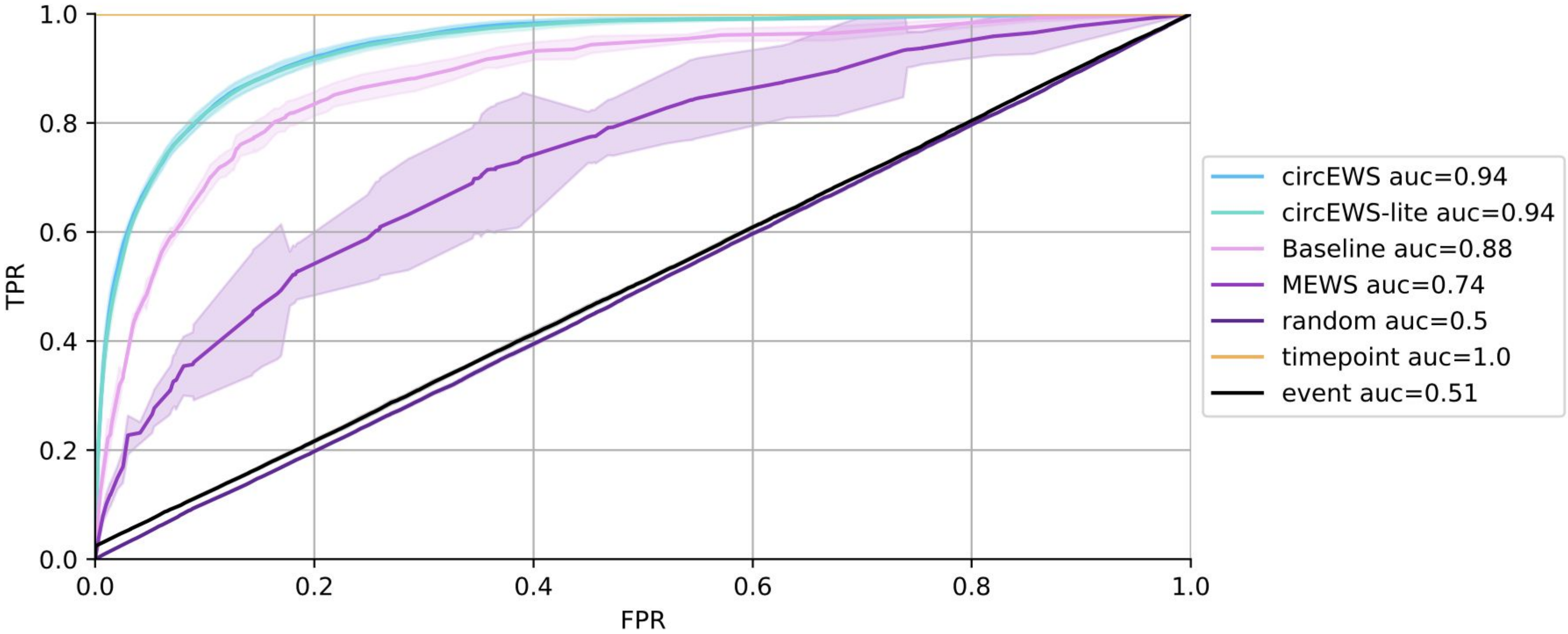
event

random

It is easy to find a threshold where random has a recall of positive timepoints than random.

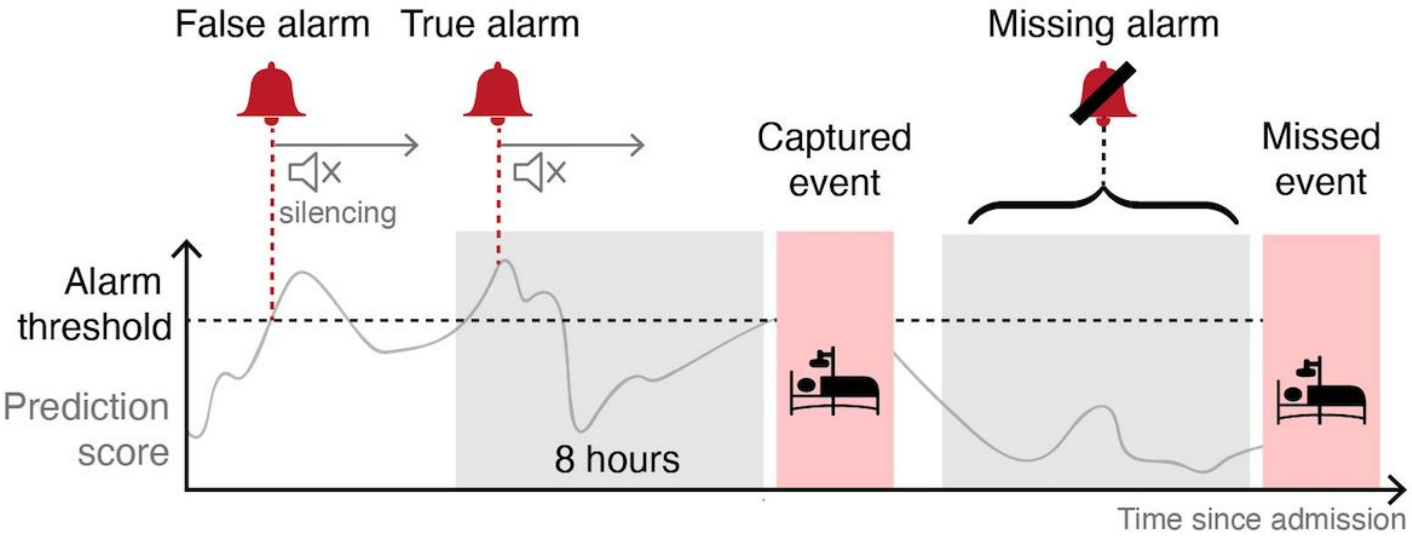
Clinical workflow requirements (such as recall of events) do not couple well with binary classification analogs

Binary classification evaluation



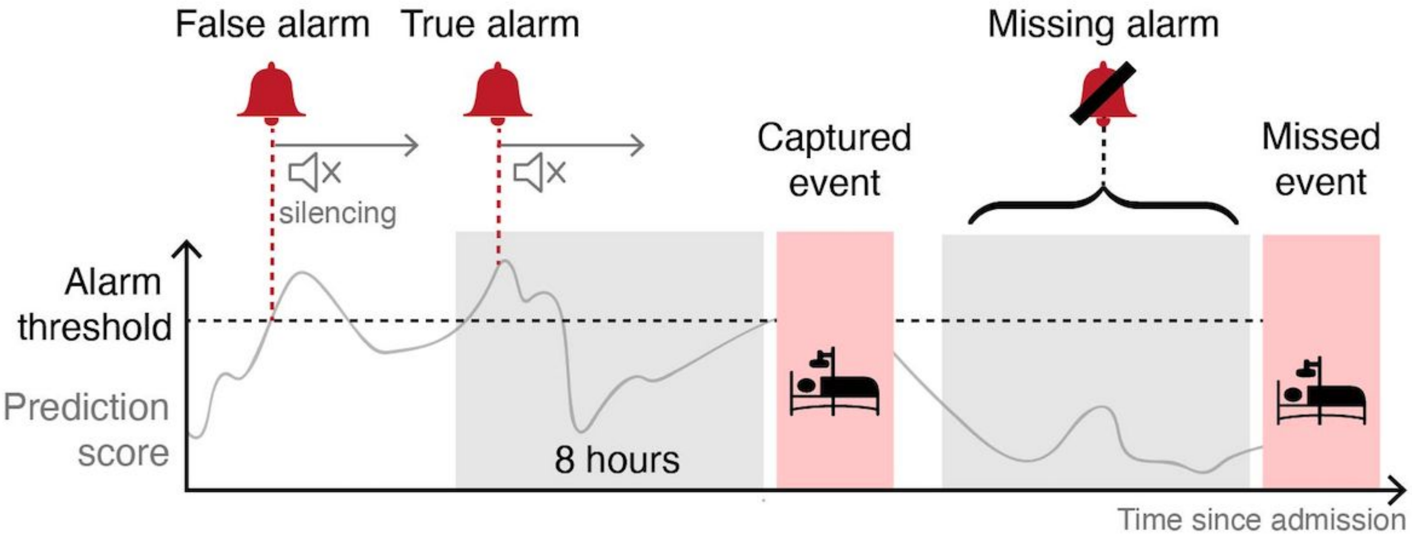
1) Event oracle performs bad due to equal weight for all positive labels.
2) What does a 0.08 difference in auROC between Baseline and circEWS mean for the clinician?

Introducing alarms



Raise an alarm if the prediction score is above the threshold and no alarm has been raised in the last 30 min

Event-based evaluation

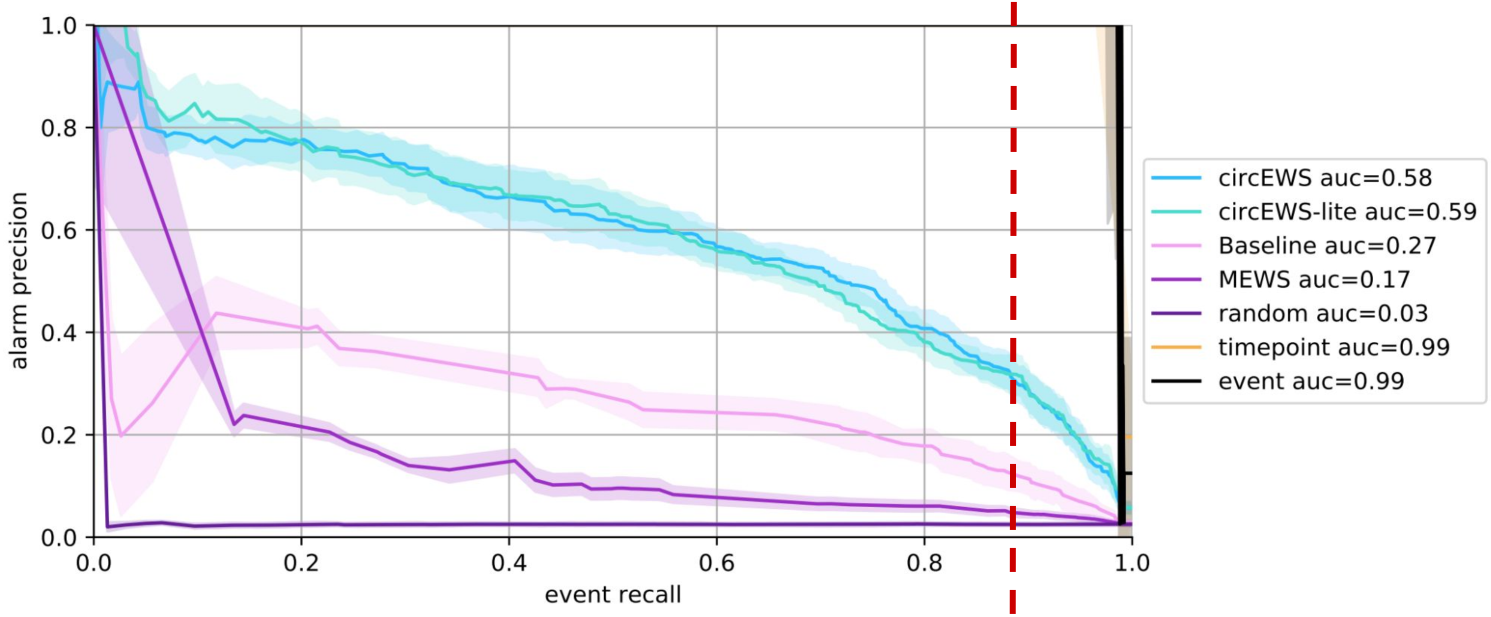


$$\textbf{Alarm}\text{ precision} = \frac{\#\text{ true alarms}}{\#\text{ total alarms}}$$

$$\textbf{Event}\text{ recall} = \frac{\#\text{ captured events}}{\#\text{ total events}}$$

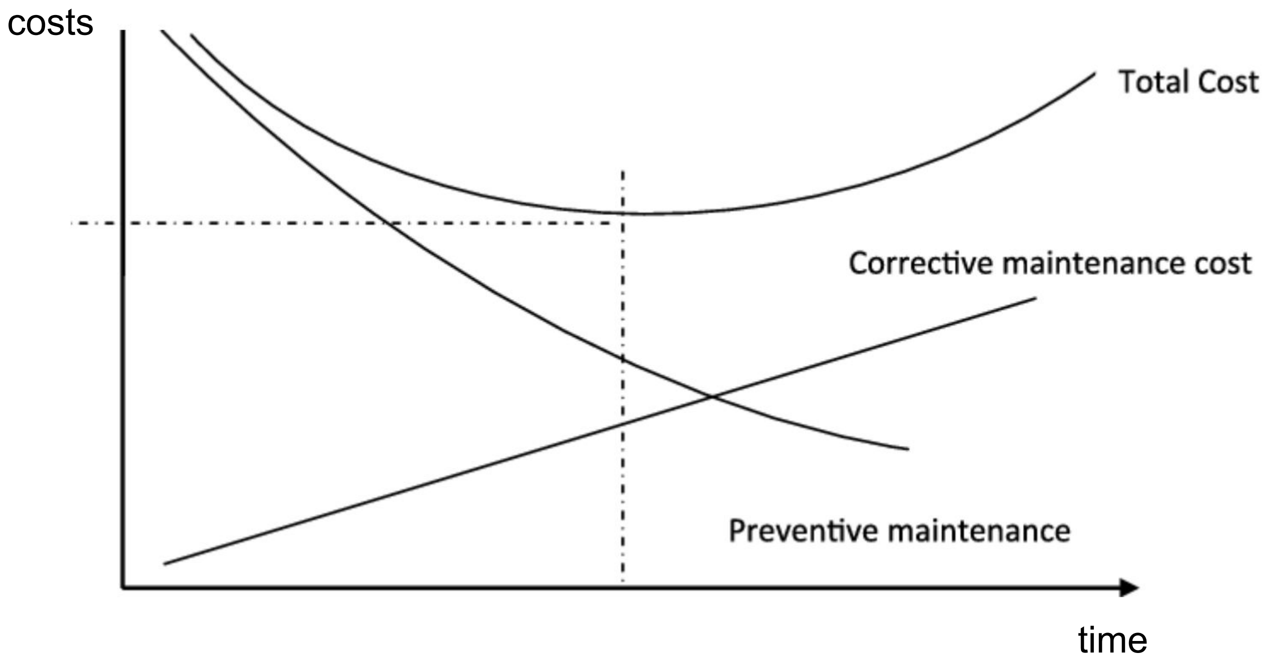
Hyland, S. L., Faltys, M., Hüser, M., Lyu, X., Gumbsch, T., Esteban, C., ... & Merz, T. M. (2020). Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature medicine*, *26*

Event-based evaluation



Is it worth reacting to alarms from circEWS?
What is the timing of the true/false alarms?

Maintenance optimization



Dawotola, Alex W., et al. "Risk-based maintenance of a cross-country petroleum pipeline system." *Journal of pipeline systems engineering and practice* 4.3 (2013): 141-148.

Maintenance policy evaluation

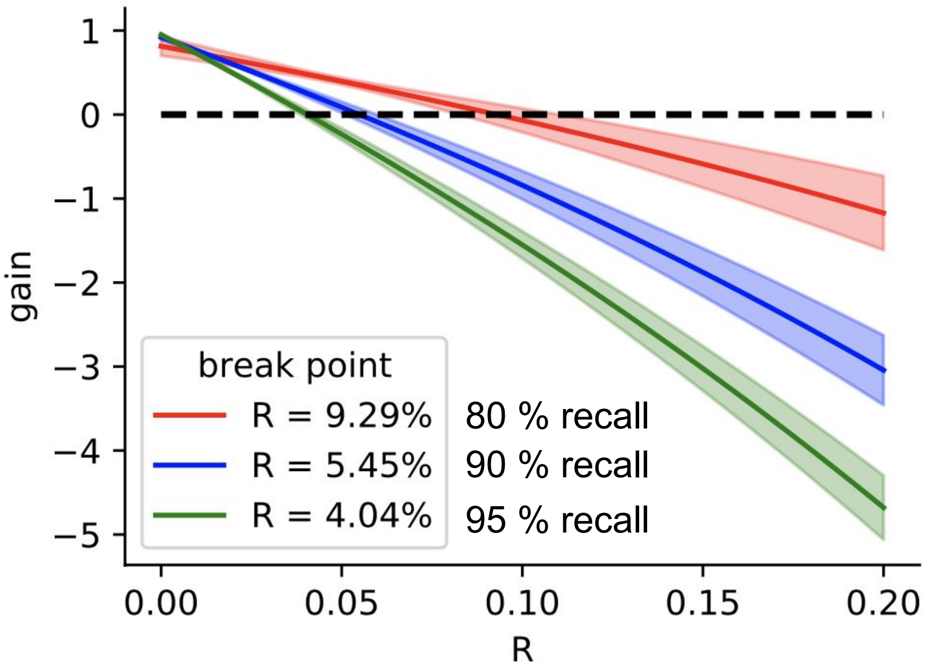
We define

1) the cost associated to reacting to an alarm (R, preventive maintenance) and
2) missing an event (1 − R, corrective maintenance).

The break point is the cost ratio R that solves

$$1 = \frac{R \cdot \#\text{total alarms}}{(1 - R) \cdot \#\text{total events}}$$

at which reacting upon the early warning system compared to when ignoring it yields the same costs.



Alaswad, S., & Xiang, Y. (2017). A review on condition-based maintenance optimization models for stochastically deteriorating system. *Reliability engineering & system safety*, *157*, 54-63.

ETH zürich

Maintenance policy evaluation

CircEWS(-lite) is 3/10 times more robust towards alarm fatigue compared to the Baseline/MEWS.

Unaccounted effects:

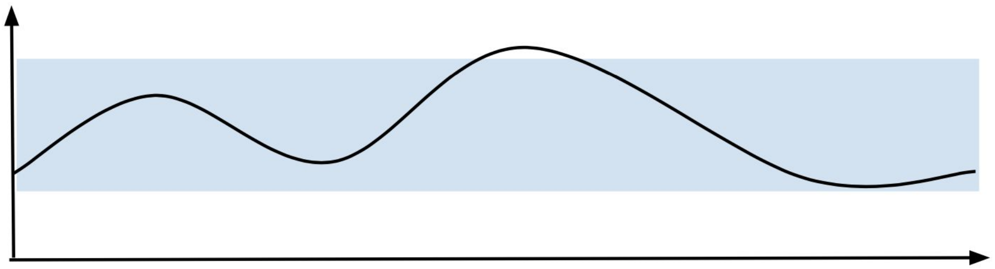
- Varying R (false alarm vs true alarm)
- Feedback loops (repeated alarms)
- Costs are intractable (time, money, resources, life satisfaction,...)

break point at recall

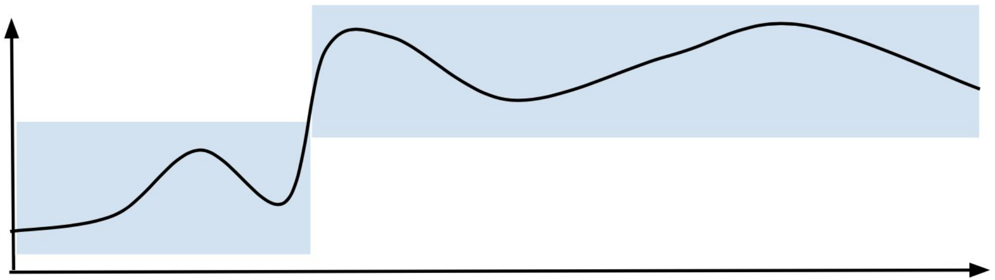
method	**R**	
	80 %	90 %
circEWS	9.14 % ± 3.52	5.5 % ± 0.87
circEWS-lite	8.5 % ± 2.17	5.16 % ± 0.68
Baseline	2.66 % ± 0.64	1.91 % ± 0.24
MEWS	0.78 % ± 0.15	0.39 % ± 0.0
random	1.16 % ± 0.19	0.78 % ± 0.19
timepoint	15.33 % ± 0.47	15.06 % ± 0.48
event	48.49 % ± 0.89	48.49 % ± 0.92

Alaswad, S., & Xiang, Y. (2017). A review on condition-based maintenance optimization models for stochastically deteriorating system. *Reliability engineering & system safety*, *157*, 54-63.

Control chart evaluation



How long does it take for a process to seem out of control just by chance?
Average run length



How many measurements are required on average to detect a shift in the underlying probability distribution
Average time to signal

Montgomery, D. C. (2020). *Introduction to statistical quality control*. John Wiley & Sons.

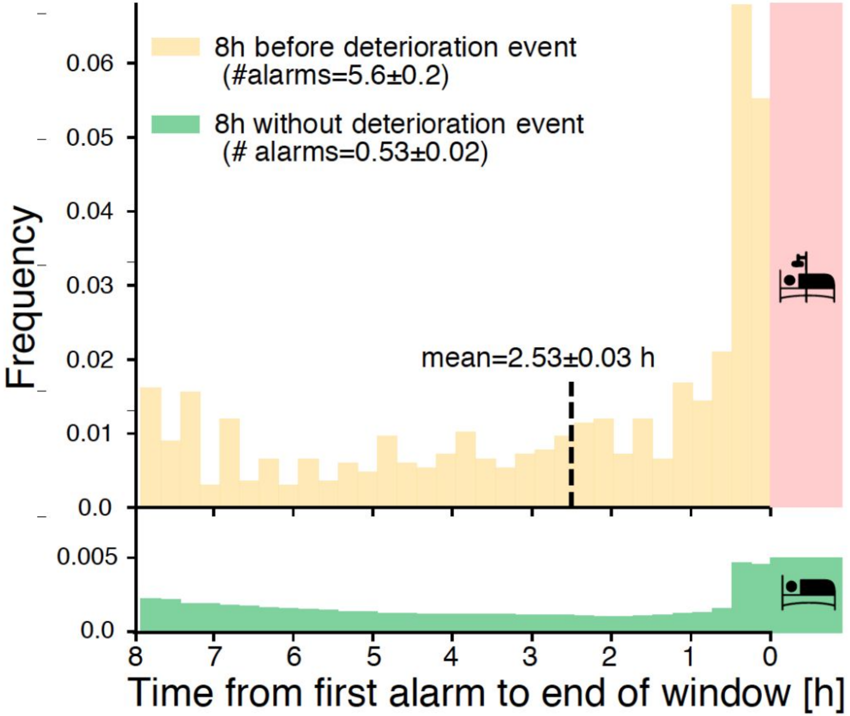
Timeliness

True alarms:

The majority of the alarms from circEWS arrive one hour before deterioration.

False alarms:

The average number of alarms in stable regions in 0.5.



Control chart evaluation

Modify control chart evaluation for event prediction

ARL0: time from one false detection to the next false detection
ARL1: time between true alarm and event

method	ARL0	
	80 %	90 %
circEWS	1h14min ± 0h11min	1h15min ± 0h2min
circEWS-lite	1h14min ± 0h9min	1h18min ± 0h1min
Baseline	1h4min ± 0h4min	1h1min ± 0h5min
MEWS	0h44min ± 0h7min	0h35min ± 0h0min
random	1h12min ± 0h5min	0h52min ± 0h3min
timepoint	nan	nan
event	nan	nan

method	ARL1	
	80 %	90 %
circEWS	2h31min ± 0h12min	2h40min ± 0h8min
circEWS-lite	2h33min ± 0h11min	2h42min ± 0h9min
Baseline	2h52min ± 0h7min	2h56min ± 0h5min
MEWS	3h0min ± 0h4min	3h1min ± 0h4min
random	3h4min ± 0h4min	3h3min ± 0h4min
timepoint	3h3min ± 0h7min	3h3min ± 0h7min
event	0h22min ± 0h2min	0h26min ± 0h2min

Montgomery, D. C. (2020). *Introduction to statistical quality control*. John Wiley & Sons.

Control chart evaluation

Seemingly in contrast to 0.5 alarms per 8h

Interpretation: Most false alarms come from intervals that also contain invalid timepoints.

method	ARL0	
	80 %	90 %
circEWS	1h14min ± 0h11min	1h15min ± 0h2min
circEWS-lite	1h14min ± 0h9min	1h18min ± 0h1min
Baseline	1h4min ± 0h4min	1h1min ± 0h5min
MEWS	0h44min ± 0h7min	0h35min ± 0h0min
random	1h12min ± 0h5min	0h52min ± 0h3min
timepoint	nan	nan
event	nan	nan

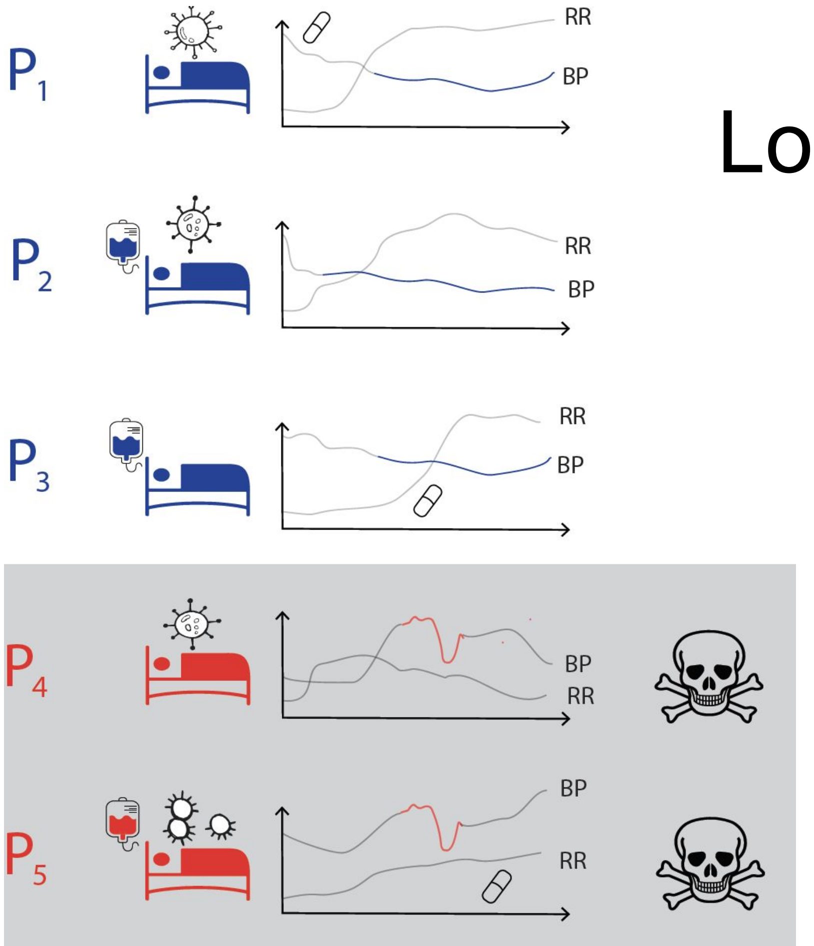
method	ARL1	
	80 %	90 %
circEWS	2h31min ± 0h12min	2h40min ± 0h8min
circEWS-lite	2h33min ± 0h11min	2h42min ± 0h9min
Baseline	2h52min ± 0h7min	2h56min ± 0h5min
MEWS	3h0min ± 0h4min	3h1min ± 0h4min
random	3h4min ± 0h4min	3h3min ± 0h4min
timepoint	3h3min ± 0h7min	3h3min ± 0h7min
event	0h22min ± 0h2min	0h26min ± 0h2min

Montgomery, D. C. (2020). *Introduction to statistical quality control*. John Wiley & Sons.

Summary evaluation

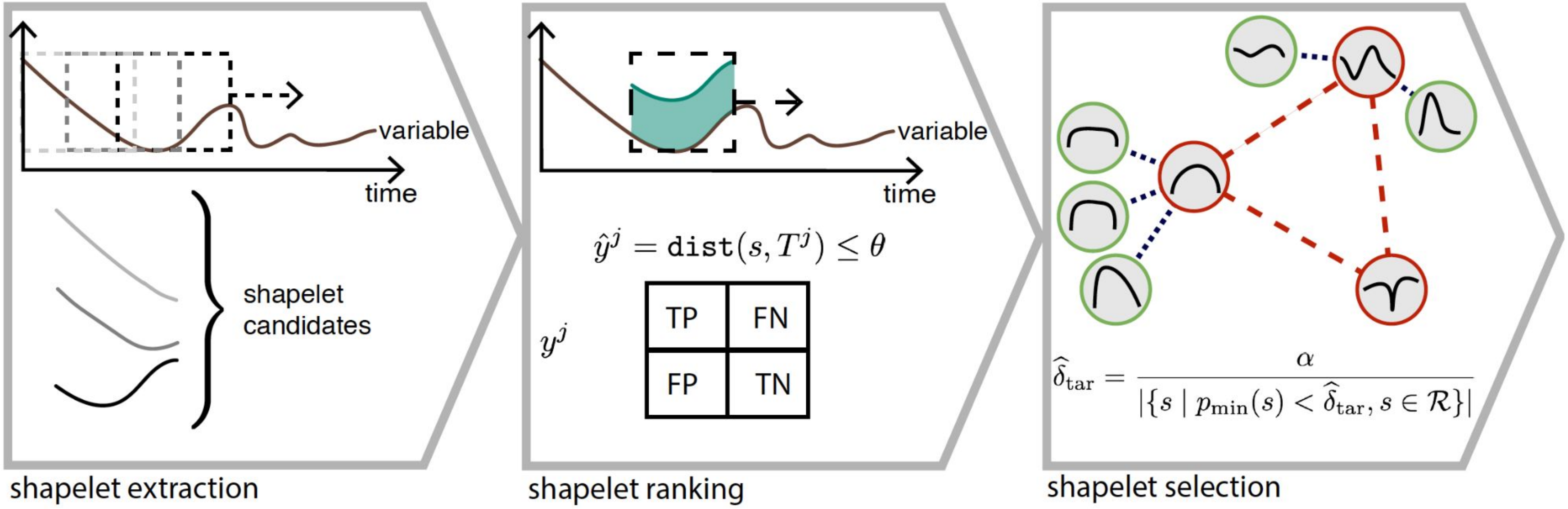
evaluation method	benefits	limitations
timepoint binary classification	readily accessible for binary classifiers reclassification assessment	meaningless for the practitioner
event-based binary classification	use-case precision and recall related to ML domain	no timeliness information user interaction inaccessible
average run length	quantification of timeliness	unable to parse invalid regions undefined for event prediction
maintenance policy	quantification of user interaction	absolute costs intractable

Part 3: Longitudinal biomarkers as features



1) Extract time series subsequences as biomarkers (so-called shapelets).
2) Use distance to shapelets at different time horizons as features for prognosis of cf

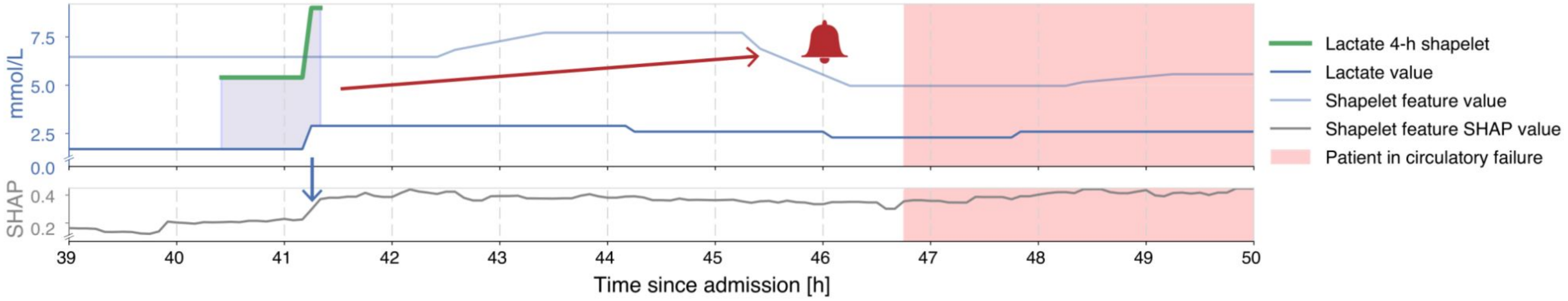
Representative shapelet mining



Gumbsch, T., Bock, C., Moor, M., Rieck, B., & Borgwardt, K. (2020). Enhancing statistical power in temporal biomarker discovery through representative shapelet mining. *Bioinformatics*, *36*

Bock, C., Gumbsch, T., Moor, M., Rieck, B., Roqueiro, D., & Borgwardt, K. (2018). Association mapping in biomedical time series via statistically significant shapelet mining. *Bioinformatics*, *34*.

Shapelet features



The lactate shapelet (in green) is most important feature based on the mean absolute SHAP value.

If circEWS encounters a region of uncertainty about the prognosis of a patient, the system reminds itself of the evidence four hours later and compares to the evidence at that later point. If at both timepoints, circulation failure is likely, an alarm is raised.

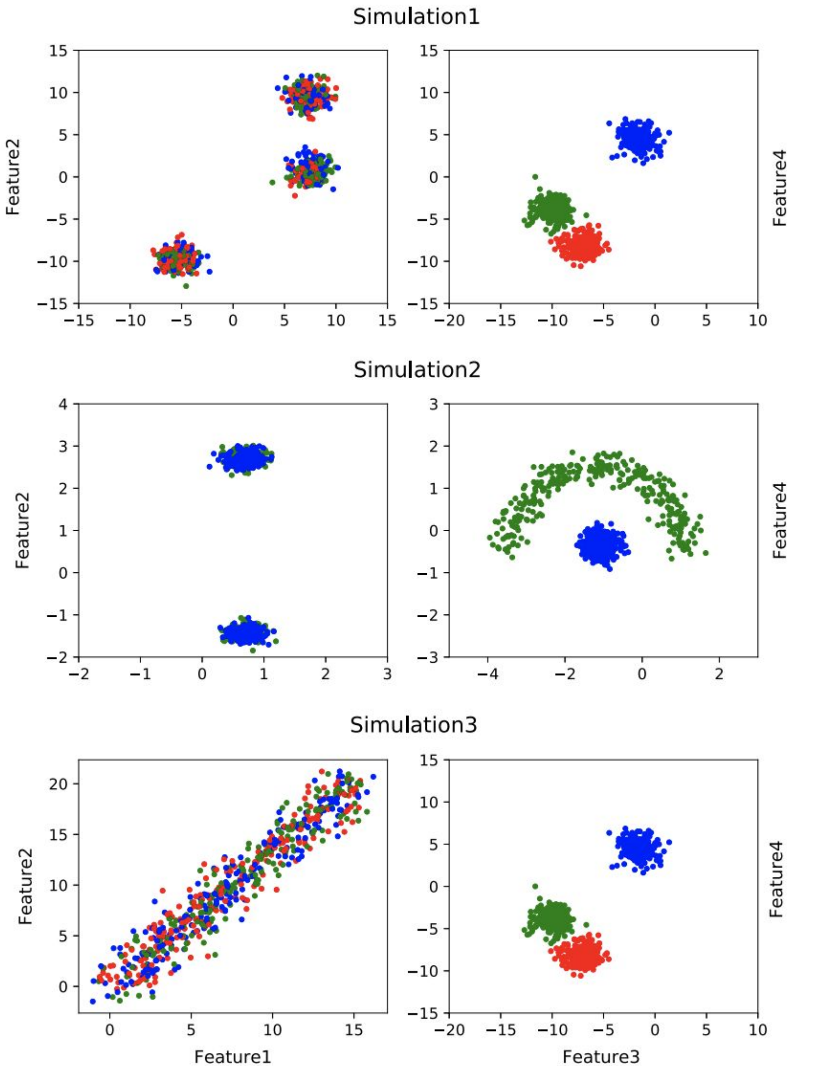
Gumbsch, T., Bock, C., Moor, M., Rieck, B., & Borgwardt, K. (2020). Enhancing statistical power in temporal biomarker discovery through representative shapelet mining. *Bioinformatics*, *36*

Part 4: Conditional clustering

Clustering that is orthogonal to given clustering.

Cluster tumor cells given the tissue type.

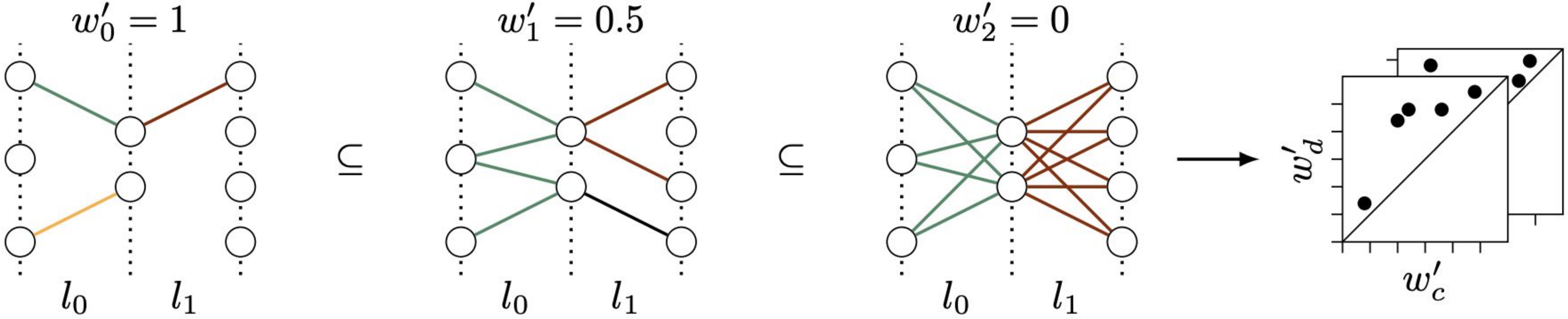
He, X., Gumbsch, T., Roqueiro, D., & Borgwardt, K. (2020). Kernel conditional clustering and kernel conditional semi-supervised learning. *Knowledge and information systems*, *62*(3), 899-925.



Part 5: Neural persistence

Topology for analyzing the state of a neural network.

Application to early stopping: Stop training if there are ‘holes’



Rieck, B., Togninalli, M., Bock, C., Moor, M., Horn, M., Gumbsch, T., & Borgwardt, K. (2018). Neural persistence: A complexity measure for deep neural networks using algebraic topology. *ICRL 2019*

Thank you

Outlook: No alarms

Relate predictions to events **without alarms**:

Prediction score gives segmentation P(t) where t is a threshold

Event definition gives segmentation E.

Given a maintenance window (where errors are allowed), compare MSE of segmentations from different predictors at optimal t.

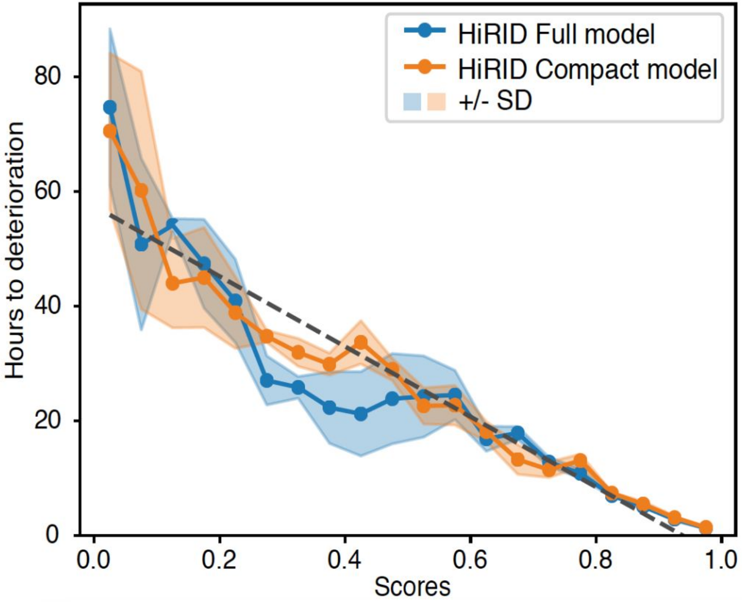
Reclassification analysis

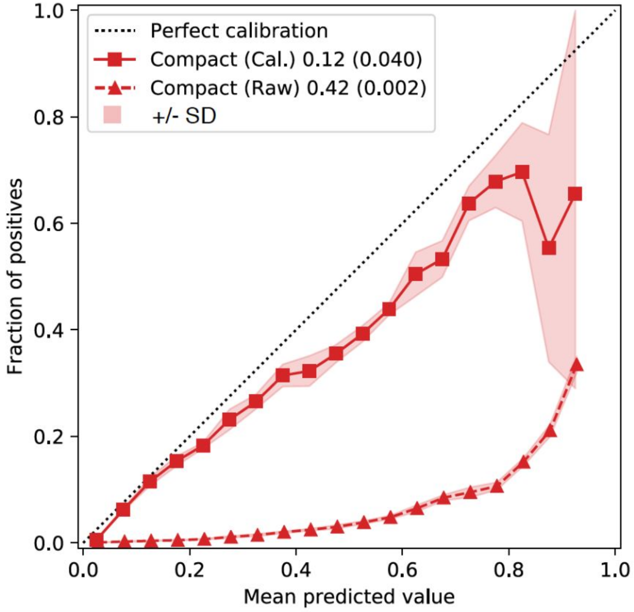
We compare circEWS to the Baseline at 0.9 recall. The table shows the fraction of reclassified timepoints.

The total fraction of reclassified timepoints is NRI=0.11, showing that the Baseline and circEWS models classify timepoints differently.

Cases	FNR circEWS	TPR circEWS
FNR Baseline	0.14 (0.11)	0.57 (0.18)
TPR Baseline	0.04 (0.13)	0.25 (0.06)
Controls	**TNR circEWS**	**FPR circEWS**
TNR Baseline	0.79 (0.06)	0.08 (0.18)
FPR Baseline	0.10 (0.17)	0.03 (0.04)

Calibration





CircEWS and CircEWS-lite are well calibrated in terms of time to failure

CircEWS-lite is poorly calibrated in terms of positive label prevalence. The calibration can be improved using isotonic regression.